AP Statistics Chapter 3 – Examining Relationships

3.1: Scatterplots and Correlation

Explanatory and Response Variables

A **response variable** measures an outcome of a study. An **explanatory variable** attempts to explain the observed outcomes. The explanatory variable is sometimes referred to as the ***independent*** variable and is typically symbolized by the variable ***x***. The response variable is sometimes referred to as the ***dependent*** variable and is typically symbolized by the variable ***y***.

Scatterplot

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of the explanatory variable appear on the horizontal axis, and the values of the response variable appear on the vertical axis. If there is no clear explanatory/response relationship between the two variables, then either variable can be placed on either axis. Each individual in the data set appears as a single point in the plot fixed by the values of both variables for that individual.

Examining a Scatterplot

In any graph of data, look for patterns and deviations from the pattern. Describe the overall **pattern** of a scatterplot by the **form**, **direction** and **strength** of the relationship.

- **Form** can be described as **linear** or **curved**.
- **Direction** can be described as **positive** or **negative** or **neither**.
- **Strength** can be described as **weak**, **moderate** or **strong**.

A **deviation** from the overall pattern of a scatterplot is called an **outlier**.

Association

- Two variables are **positively associated** if as one increases the other increases.
- Two variables are **negatively associated** if as one increases the other decreases.

Correlation

Correlation measures the strength and direction of the relationship between two quantitative variables. Correlation is usually represented by the letter ***r***.

Facts about Correlation

1. When calculating correlation, it makes no difference which variable is x and which is y.
2. Correlation is only calculated for quantitative variables, not categorical.
3. The value of r does not change if the units of x and/or y are changed.
4. Positive r indicates a positive association between x and y. Negative r indicates a negative association.
5. Correlation is always a number between -1 and +1. Values close to +1 or -1 indicate that the points lie close to a line. The extreme values of +1 and -1 are only achieved when the points are perfectly linear.
6. Correlation measures the strength of a linear relationship between two variables, not curved relationships.
7. Correlation, like the mean and standard deviation, is nonresistant. Recall that this means that it is greatly affected by outliers.

3.2: Least-Squares Regression

Regression Line

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. The line is often to predict values of y for given values of x. Regression, unlike correlation, requires an explanatory/response relationship. In other words, when x and y are reversed, the regression line changes. Recall that correlation is the same no matter which variable is x and which is y.

Least-Squares Regression Line

The **least-squares regression line** is the line that makes the sum of the squares of the vertical distances from the data points to the line as small as possible.

Equation of the Least-Squares Regression Line

To find the equation of the regression line in the form $y = a + bx$, where a is the y-intercept and b is the slope, use the following equations:

$$b = r\frac{s_y}{s_x} \quad \text{and} \quad a = \overline{y} - b\overline{x}$$

The Role of r-squared (Coefficient of Determination)

The square of the correlation coefficient, or **r-squared**, represents the percentage of the change in the y-variable that can be attributed to its relationship with the x-variable. So if r-squared for the regression between x and y is .73, we can say that x accounts for 73% of the variation in y.

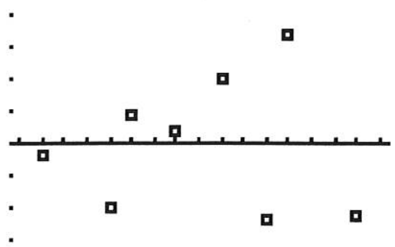
Residuals

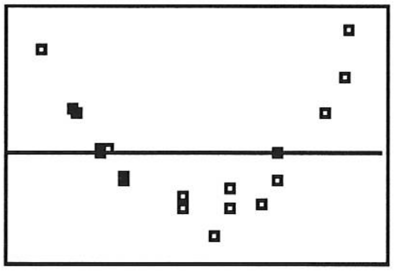
A residual is the difference between an observed value of y and the value predicted by the regression line. That is, residual = actual y - predicted y.

Residual Plot

A residual plot is a scatterplot of each x-value and its residual value. The residual plot is used to determine whether a linear equation is a good model for a set of data, as follows:

- If the residual plot exhibits randomness, then a line is a good model for the data (see left)
- If the residual plot exhibits a pattern, then a line is NOT a good model for the data (right)





Outliers and Influential Points

A point that lies outside the overall pattern of the other observations is considered an **outlier**. If the removal of such a point has a large effect on the correlation and/or regression, that point is considered an **influential point**.

AP Statistics - Chapter 3 MC Study Guide

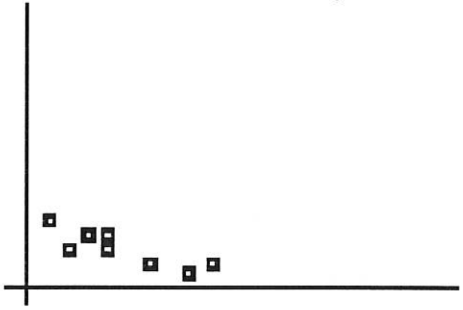
Multiple Choice

Identify the choice that best completes the statement or answers the question.

□ 1. In a statistics course, a linear regression equation was computed to predict the final exam score from the score on the first test. The equation was $y = 10 + .9x$ where y is the final exam score and x is the score on the first test. Carla scored 95 on the first test. On the final exam, Carla scored 98. What is the value of her residual?

a. 98
b. 2.5
c. –2.5
d. 0
e. None of the above

□ 2. In the scatterplot below, if each x-value were decreased by one unit and the y-values remained the same, then the correlation *r* would



a. Decrease by 1 unit
b. Decease slightly
c. Increase slightly
d. Stay the same
e. Can't tell without knowing the data values

□ 3. In regression, the residuals are which of the following?

a. Those factors unexplained by the data
b. The difference between the observed responses and the values predicted by the regression line
c. Those data points which were recorded after the formal investigation was completed
d. Possible models unexplored by the investigator
e. None of the above

□ 4. Which of the following statements are true?

I. Correlation and regression require explanatory and response variables.

II. Scatterplots require that both variables be quantitative.

III. Every least-square regression line passes through $(\bar{x}, \bar{y})$.

a. I and II only
b. I and III only
c. II and III only
d. I, II, and III
e. None of the above

□ 5. Suppose the following information was collected, where X = diameter of tree trunk in inches, and Y = tree height in feet.

X	4	2	8	6	10	6
Y	8	4	18	22	30	8

If the LSRL equation is y = –3.6 + 3.1x, what is your estimate of the average height of all trees having a trunk diameter of 7 inches?

a. 18.1
b. 19.1
c. 20.1
d. 21.1
e. 22.1

□ 6. Suppose we fit the least squares regression line to a set of data. What is true if a plot of the residuals shows a curved pattern?

a. A straight line is not a good model for the data.
b. The correlation must be 0.
c. The correlation must be positive.
d. Outliers must be present.
e. The LSRL might or might not be a good model for the data, depending on the extent of the curve.

□ 7. Which of the following are resistant?

a. Least squares regression line
b. Correlation coefficient
c. Both the least squares line and the correlation coefficient
d. Neither the least squares line nor the correlation coefficient
e. It depends

□ 8. A copy machine dealer has data on the number x of copy machines at each of 89 customer locations and the number y of service calls in a month at each location. Summary calculations give $\bar{x}$ = 8.4, S_x = 2.1, $\bar{y}$ = 14.2, S_y = 3.8, and r = 0.86. What is the slope of the least squares regression line of number of service calls on number of copiers?

a. 0.86
b. 1.56
c. 0.48
d. None of these
e. Can't tell from the information given

□ 9. There is a linear relationship between the number of chirps made by the striped ground cricket and the air temperature. A least squares fit of some data collected by a biologist gives the model $\hat{y} = 25.2 + 3.3x$, $9 < x < 25$, where x is the number of chirps per minute and $\hat{y}$ is the estimated temperature in degrees Fahrenheit. What is the estimated increase in temperature that corresponds to an increase in 5 chirps per minute?

a. 3.3°F
b. 16.5°F
c. 25.2°F
d. 28.5°F
e. 41.7°F

□ 10. A set of data relates the amount of annual salary raise and the performance rating. The least squares regression equation is $\hat{y} = 1{,}400 + 2{,}000x$ where y is the estimated raise and x is the performance rating. Which of the following statements is *not* correct?

a. For each increase of one point in performance rating, the raise will increase on average by $2,000.
b. This equation produces predicted raises with an average error of 0.
c. A rating of 0 will yield a predicted raise of $1,400.
d. The correlation for the data is positive.
e. All of the above are true.

AP Statistics Name:

Directions: *Work on these sheets.*

Part 1: Multiple Choice *Circle the letter corresponding the best answer.*

1. A study found correlation $r = 0.61$ between the sex of a worker and his or her income. You conclude that
(a) women earn more than men on the average.
(b) women earn less than men on average.
(c) an arithmetic mistake was made; this is not a possible value of r.
(d) this is nonsense because r makes no sense here.
(e) the correlation should have been $r = -0.61$.

2. A copy machine dealer has data on the number x of copy machines at each of 89 customer locations and the number y of service calls in a month at each location. Summary calculations give $\bar{x} = 8.4$, $s_x = 2.1$, $\bar{y} = 14.2$, $s_y = 3.8$, and $r = 0.86$. What is the slope of the least-squares regression line of number of service calls on number of copiers?
(a) 0.86
(b) 1.56
(c) 0.48
(d) None of these
(e) Can't tell from the information given

3. In the setting of the previous problem, about what percent of the variation in the number of service calls is explained by the linear relation between number of service calls and number of machines?
(a) 86%
(b) 93%
(c) 74%
(d) None of these
(e) Can't tell from the information given

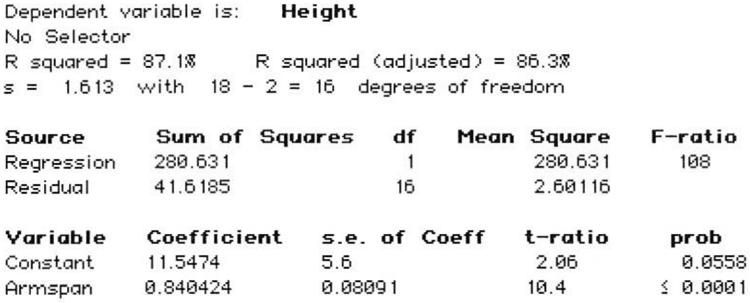
4. If data set A of (x, y) data has correlation coefficient $r = 0.65$, and a second data set B has correlation $r = -0.65$, then
(a) the points in A exhibit a stronger linear association than B.
(b) the points in B exhibit a stronger linear association than A.
(c) neither A nor B has a stronger linear association.
(d) you can't tell which data set has a stronger linear association without seeing the data or seeing the scatterplots.
(e) a mistake has been made—r cannot be negative.

5. There is a linear relationship between the number of chirps made by the striped ground cricket and the air temperature. A least-squares fit of some data collected by a biologist gives the model $\hat{y} = 25.2 + 3.3x,\ \ 9 < x < 25$, where x is the number of chirps per minute and $\hat{y}$ is the estimated temperature in degrees Fahrenheit. What is the estimated increase in temperature that corresponds to an increase of 5 chirps per minute?
(a) 3.3°F (b) 16.5°F (c) 25.2°F (d) 28.5°F (e) 41.7°F

6. Which of the following relationships is most likely to result in a strong negative correlation?
(a) The number of people showering in a college dorm and the water pressure in each shower.
(b) The outdoor temperature and the number of fans running in non-air-conditioned dorm rooms.
(c) The comfort rating of a mattress and the number of hours of uninterrupted sleep obtained.
(d) The price of a home and its square footage.
(e) The fuel efficiency of a car (miles per gallon) and its speed.

7. A set of data relates the amount of annual salary raise and the performance rating. The least squares regression equation is $\hat{y} = 1400 + 2000x$ where y is the raise amount and x is the performance rating. Which of statements (a) to (d) is *not* correct?
(a) For each increase of one point in performance rating, the raise will increase on average by $2000.
(b) This equation produces predicted raises with an average error of 0.
(c) A rating of 0 will yield a predicted raise of $1400.
(d) The correlation between salary raise and performance rating is positive.
(e) All of the above are true.

8. Leonardo da Vinci, the renowned painter, speculated that an ideal human would have an armspan (distance from outstretched fingertip of left hand to outstretched fingertip of right hand) that was equal to his height. The following computer regression printout shows the results of a least-squares regression on height and armspan, in inches, for a sample of 18 high school students.



```
Dependent variable is:    Height
No Selector
R squared = 87.1%      R squared (adjusted) = 86.3%
s =   1.613  with   18 - 2 = 16   degrees of freedom

Source        Sum of Squares    df    Mean Square    F-ratio
Regression    280.631            1        280.631      108
Residual      41.6185           16        2.60116

Variable     Coefficient    s.e. of Coeff    t-ratio       prob
Constant     11.5474        5.6               2.06         0.0558
Armspan      0.840424       0.08091          10.4        ≤ 0.0001
```

Which of the following statements is *false*?
(a) This least-squares regression model would make a prediction that is 1.63 inches higher than da Vinci projected for a 62-inch tall student.
(b) One of the students in the sample had a height of 70.5 inches and an armspan of 68 inches. The residual for this student is 1.83 inches.
(c) Da Vinci's projection is lower than the prediction that this least-squares line will make for any height.
(d) For every one-inch increase in armspan, the regression model predicts about a 0.84-inch increase in height.
(e) For a student 66 inches tall, our model would predict an armspan of about 67 inches.

Part 2. Free Response

Answer completely, but be concise. Show your work.

Joey appears to be growing slowly as a toddler. His height between 18 and 30 months of age increases as follows:

Age (months)	Observed height (cm)	Predicted height	Residual
18	76.5		-0.08
21	78.7	79.09	
24	82.0	81.6	0.4
27	84.8	84.11	
30	86.0		-0.62

The least-squares regression line fitted to this data has equation

HEIGHT = 61.5 + 0.837 AGE

9. Finish filling in the table above.

10. Sketch a residual plot on the axes provided.

11. Based on your residual plot, would you describe Joey's growth pattern from 18 to 30 months as linear? Explain.

12. According to the least-squares principle, which of the lines below provides the best fit for the data shown in the scatterplot? Justify your answer.

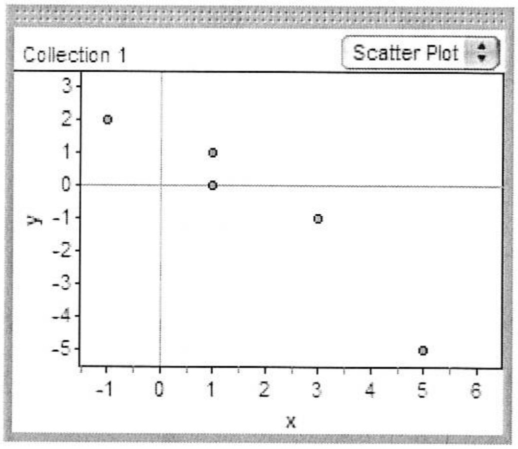
(a) $y = 2 - x$

(b) $y = 1.5 - x$

(c) $y = 1 - x$

(d) $y = 3 - 2x$

(e) $y = 3 - 1.5x$



13. Anthropologists must often estimate from human remains how tall the person was when alive. Carla is studying how overall height can be predicted from the length of a leg bone in a group of 36 living males. The data show that the bone lengths have mean 45.9 cm and standard deviation 4.2 cm, the overall heights have mean 172.7 cm and standard deviation 8.14 cm, and the correlation between bone length and height is 0.914.

(a) Determine the equation of the least-squares regression line of height on bone length. Show your work.

(b) Interpret the correlation in the context of this problem.

14. In general, is correlation a resistant measure of association? ________ Explain briefly or give a simple example to illustrate.